Effective Sample Size Estimation for a Mechanical Ventilation Trial Through Monte-Carlo Simulation: Length of Mechanical Ventilation and Ventilation Free Days.

Sophie Morton [a,*]	sophie.morton@pg.canterbury.ac.nz
Yeong Shiong Chiew [a,1]	chiew.yeong.shiong@monash.edu
Christopher Pretty [a]	chris.pretty@canterbury.ac.nz
Elena Moltchanova [b]	elena.moltchanova@canterbury.ac.nz
Carl Scarrott [b]	carl.scarrott@canterbury.ac.nz
Daniel Redmond [a]	daniel.redmond@pg.canterbury.ac.nz
Geoffrey M Shaw [c]	geoff.shaw@cdhb.health.nz
J. Geoffrey Chase [a]	geoff.chase@canterbury.ac.nz

---

[a] Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand

[b] School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

[c] Intensive Care Unit, Christchurch Hospital, Christchurch, New Zealand

* Corresponding author

[1] Present address: Malaysia School of Engineering, Monash University Sunway Malaysia

---

Keywords: Statistical significance, mechanical ventilation, Monte Carlo methods, randomised control trials, outcome metrics, ventilator free days

---

Abbreviations Used:

ARDS	Acute Respiratory Distress Syndrome
ICU	Intensive Care Unit
KS-test	Kolmogorov-Smirnov Test
LoMV	Length of Mechanical Ventilation
LoMV-28	Length of Mechanical Ventilation - 28 Days
MV	Mechanical Ventilation
RCT	Randomised Control Trial
RS-test	Wilcoxon-Ranksum Test
VFD	Ventilator Free Days

Abstract

Randomised control trials (RCTs) have sought to seek to improve mechanical ventilation (MV) treatment. However, few RCTs to date have shown clinical significance. It is hypothesised that aside from effective treatment, the outcome metrics and sample sizes of the trial also affect the significance, and thus impact trial design.

In this study, a Monte-Carlo simulation method is developed and used to investigate several outcome metrics of MV treatment, including 1) Length of Mechanical Ventilation (LoMV); 2) ventilator free days (VFD); and 3) LoMV-28. In addition, it investigates the impact of imposing clinically relevant exclusion criteria on study power to enable better design for significance. Data from invasively ventilated patients from a single intensive care unit were used in this analysis to demonstrate the method.

Use of LoMV as an outcome metric found that 160 patients/arm were required to reach 80% power with a clinically expected intervention difference of 25% LoMV if clinically relevant exclusion criteria are applied to the cohort, but 400 patients/arm if they are not. However, 130 patients/arm would be required for the same statistical significance at the same intervention difference if VFD is used.

A Monte-Carlo simulation approach using local cohort data combined with objective patient selection criteria can yield better design of ventilation studies to desired power and significance, but with fewer patients per arm than traditional trial design methods, which in turn reduces patient risk. Use of outcome metrics, such as VFD should be used when there is also expected to be a difference in mortality between the two cohorts. Finally, the non-

parametric approach taken is readily generalisable to a range of trial types where outcome data is similarly skewed.

1. Introduction

Mechanical Ventilation is a core intensive care therapy for patients suffering from respiratory failure or acute respiratory distress syndrome (ARDS) [1]. While it is a relatively straightforward treatment, optimising mechanical ventilation without causing damage to the lung is complex in practice. A range of randomised control trials (RCTs) have been carried out to assess methods of improving patient MV care. However, many have had non-significant findings [2–5], and the field remains uninformed about consistent action that might improve outcomes.

Respiratory failure is often a secondary symptom from a range of diseases, many causing lung damage that is mixed in effect and severity [6]. Thus, the generalised treatment proposed in some RCTs may not provide the best possible treatment for all patient types. In addition, insignificant RCT results may also be partly due to difficulty in determining the efficacy of mechanical ventilation therapy. Aside from patient mortality, other metrics used to assess the quality of mechanical ventilation treatment include cardiopulmonary and haemodynamic responses, patients' physiological or acuity scores, and patients' ventilator dependency such as length of mechanical ventilation (LoMV) and ventilator free days (VFDs). However, all these metrics have limitations.

LoMV or VFD are the two most common metrics that were used to assess MV efficacy. These metrics consider patient ventilator dependency and how early patients are weaned from the ventilator along with the mortality rate for the cohort [7]. They also assess the economic impact, as ventilator dependency is associated with higher cost [8].

For a clinical trial to be successful, it must have both useful results and statistical significance [9]. While a trial may have useful clinical results, it is unable to make a meaningful statement without sufficient statistical significance or 'power'. Thus, determining the necessary effective trial sample size to reach a sufficient power is critical. Table 1 shows a range of mechanical ventilation RCTs that use LoMV or VFD as one of their outcome metrics [2–5,10–13]. These studies ranged in size from 70-2300 patients, with only two able to reach a statistical significance of $p<0.05$.

Table 1: Summary of several randomised control trials assessing LoMV and VFD

Study	**No. Patient**	**Metric Used**	**Groups (Number of patient) LoMV or VFD** (in mean ± standard deviation or median [interquartile range])		**p-value**
ARDSNet [5]	861	VFD	Low Vt+ (432) 12±11	High Vt (429) 10±11	**0.0070**
ALVEOLI [2]	549	VFD	Lower PEEP# (273) 14.5±10.4	Higher PEEP (276) 13.8±10.6	**0.5000**
EXPRESS [4]	767	VFD	Minimal distension (382) 3 [0-17]	Increased recruitment (385) 7 [0-19]	**0.0400**
LOVS [3]	983	LoMV	Control (507) 10 [6-16]	Lung open (475) 10 [6-17]	**0.9200**
Meta-Analysis [14]	2299	VFD	Lower PEEP (1136) 11 [0-21]	Higher PEEP (1163) 13 [0-22]	**0.1000**
Individualised PEEP [10]	70	VFD	Control (36) 0 [0-15.75]	Intervention (34) 1 [0-18]	**0.1600**
Sedation Study [11]	113	VFD	Control (58) 18.0 [0-24.1]	No Sedation (55) 6.9 [0-20.5]	**0.0191**

When clinical significance was not found, it was often due to ineffective treatment or inability to effectively treat all patients. However, high levels of patient variability as well as insufficient sample sizes can significantly impact the ability of a clinical study to achieve significance [15,16]. In an earlier study by Chiew et al [17], it was also noted that the commonly used sample size estimation methods for a powered study [18] were not feasible for LoMV clinical

data that were heavily skewed with a very long upper tail. Thus, it is not possible to truly assess whether trial design or numbers, or trial inefficacy are the course of failure. Hence, a simulation-based method using retrospective clinical cohort data may provide a better estimation of a well-powered sample size for a desired outcome metric and patient cohort [19].

This study presents a Monte-Carlo simulation-based method to estimate sample sizes for a powered and significant RCT for a range of outcome metrics relating to ventilator dependency. The outcome metrics investigated in this study were LoMV, VFD and a modified LoMV. A case study for determining the sample sizes of a planned RCT is also presented, where patient selection criteria are also simulated to replicate the planned RCT as closely as possible [20]. Overall this study presents a non-parametric simulation based method that is readily generalisable for trial design, and presents it in terms of a sample size study design involving LoMV and VFD, their potential limitations, including a case example which also demonstrates how this method can effectively pre-test a cohort when designing the trial.

2. Methods

2.1. Sample size analysis metric

Three outcome metrics for sample size estimation were investigated: 1) Length of mechanical ventilation (LoMV); 2) Ventilation free days in 28 days (VFD) [7]; and 3) Length of mechanical ventilation within 28 days (LoMV-28). VFD and LoMV-28 are modified LoMV distributions that also include mortality information where deceased patients have 0 VFD or 28 days of LoMV. Table 2 shows a more detailed description of each outcome metric used in this study.

Table 2: Outcome metrics to be used in study

1.	**LoMV:**	The total duration of mechanical ventilation.
2.	**VFD:**	The number of days free of MV within a 28 day period. VFD is defined by [7] as: • VFD = 0: if the patient dies before 28 days • VFD = (28 – LoMV): if the patient is successfully weaned from MV within 28 days. • VFD = 0: if the patients requires MV for 28 days or more
3.	**LoMV-28:**	Length of MV within 28 days, where: • LoMV-28 = 28: if the patient dies before 28 days • LoMV-28 = LoMV: if the patient is successfully weaned from MV within 28 days • LoMV-28 = 28: if the patients required MV for 28 days or more.

2.2. Retrospective patient cohort data (Cohort A)

Retrospective data from 5176 patients admitted to the Christchurch Hospital Intensive Care Unit (ICU) from 2011 to 2014 was considered in this study. All APACHE III diagnostic codes, ICU mortality and length of mechanical ventilation (LoMV) were recorded. Of this number, 3896 (75%) patients required MV therapy and 3383 (63%) received invasive ventilation either through tracheotomy or intubation.

In this study, only patients who were invasively ventilated are considered, which is the largest possible cohort and delineated Cohort A in this study. The mean LoMV is 2.95 ± 6.50 days (median = 0.73 [IQR: 0.24 – 2.48]). The detailed patient distribution for this cohort and their corresponding LoMV and mortality distribution can be found in Appendix A, Tables A1 to A3. The description of APACHE III diagnostic codes is included in Appendix B.

2.3. Simulating realistic clinical trial cohorts (Cohort B)

Not all invasively ventilated patients may benefit from optimised MV. For example, some patient groups receive MV only for brief post-surgical periods. Hence, they would not be part of such a trial. The difficulty is that exclusion criteria in many trials are subjective and thus can add variability and unintended dimensionality to the study, affecting the potential outcome in ways not included in the study design.

Using Monte-Carlo simulation, an objective patient cohort can be created and simulated from Cohort A. This objective cohort (Cohort B) aims to capture the realistic characteristics of a patient cohort expected to be used in a planned clinical trial. Objective patient selection is enabled using the APACHE III diagnostic code to simulate the actual clinical trial inclusion and exclusion criteria. Many of these criteria have been used in prior studies [2–5,10–12], and objective criteria for all exclusions would ensure a more robust design and implementation.

The exclusion criteria typically used are listed below and the studies they are used in are referenced. These criteria include:

1. Patients who are likely to be discontinued from MV within 24 hours [10,11];
2. Patients with raised intracranial pressure [2–5,10–12];

3. Patients who have significant weakness from any neurological disease [2–5,10];
4. Patients who have asthma as the primary presenting condition, or a history of significant chronic obstructive pulmonary disease [2–5].

In this study, a sample of clinical inclusion and exclusion criteria for a randomised controlled trial is used (ANZCTR number: ACTRN12614001069640). Inclusion criteria are set to target every patient that is eligible for the study (Cohort A). The exclusion criteria are chosen based on the clinical implication that these patients may not benefit from a MV intervention, or could be harmed in some cases – as listed above.

In this study, the objective cohort (Cohort B) was established by excluding all patients under APACHE III diagnostic codes as shown in Appendix Table A4, which are relevant to the 4 main criteria typically used and listed above. The use of diagnostic codes avoids also subjective choices in both simulation and implementation, where such subjectivity is difficult to model and induces unintended variability from what might actually occur. It could also be easily and objectively implemented in a real trial, which would better ensure that the trial design and the actual study matched.

Thus, the following specific APACHE III diagnostic codes were also excluded in implementing the typical exclusion criteria listed previously:

- 206 - Chronic obstructive pulmonary disease
- 209 - Asthma
- 400 - Neurological non-operative
- 601 - Head trauma with or without multi trauma
- 604 - Multi trauma with spinal injury
- 605 - Isolated cervical spine injury
- 1500 - Neurological post-operative
- 1601 - Post operation patients: head trauma with or without multi trauma

- 1604 - Post operation patients: Multi trauma with spinal injury
- 1605 - Post operation patients: isolated cervical spine injury

This approach makes the criteria objective and easy to implement both in simulation and in a clinical trial. After imposing the exclusion criteria, the number of patients eligible for the study is reduced to 974 (18.8% of total patients admitted to ICU or 28.8% of patients requiring invasive MV). This cohort is denoted as Cohort B. A detailed comparison of the actual trial exclusion criteria and simulation method is shown in Appendix Table B1.

2.4. Sample size determination using Monte-Carlo simulation

A Monte-Carlo simulation was performed to determine the power of the study at a range of sample sizes. This simulation allows a range of intervention effects to be simulated, and the corresponding sample sizes required to detect the significance at a power, to be calculated. A 10,000 iteration Monte-Carlo simulation was run over the data to determine the required sample size for each arm of the study to achieve 80% power. In this study, the sample sizes for Cohort A and Cohort B with different characteristics were examined.

Both double-sided and single-sided log-normal Student t-test (t-test), Wilcoxon Rank-Sum (RS-test) and a Kolmogorov-Smirnov test (KS-test) were used for significance testing of the difference in mean and other distribution characteristics (median and variability). All simulations were performed using MATLAB [21].

The hierarchical steps followed to carry out the Monte-Carlo design analysis are outlined in Table 3. The change into VFD and LoMV-28 metrics is carried out after the LoMV difference has been imposed on the intervention group (Step 3) and before the statistical testing (Step 4).

Table 3: Steps followed in Monte Carlo Simulation

	Step	Description	Tested Group
1.	**Patient Cohorts**	• Select a patient cohort 1) Cohort A includes all invasively ventilated patients or 2) Cohort B is created from Cohort A by imposing exclusion criteria.	• Cohort A • Cohort B
2.	**Sample Size Selection**	• Randomly select patients from the patient cohort and assign each patients to a treatment group. 1) Control group or 2) Intervention group. • Patient is selected with replacement. • Various sample size of each treatment group is tested.	• Total sample size N= 100, 110, 120… 2000 patients
3.	**Difference in LoMV**	• Impose an intervention effect to simulate differences in LoMV between the two groups. • The LoMV in Intervention group is reduced by the chosen percentage. • LoMV intervention = LoMV patient × (100% -Percentage reduction) • The difference in LoMV ranges from 10 to 25% of total LoMV.	• Difference of LoMV = 10%, 15%, 20%, 25%
4.	**Metric Conversion**	• Calculation of VFD and LoMV-28 using given LoMV.	• Cohort A • Cohort B
5.	**Statistical Test**	• Perform statistical test comparing the metrics (LoMV, VFD, LoMV-28) between two groups. • Using parametric and non-parametric tests. • A p-value < 0.05 indicates that LoMV for intervention group is significantly different from control group.	• Student *t*-test (log scale) • RS test • KS test
6.	**Power Analysis**	• Each Monte-Carlo simulation iteration will generate a p-value for each statistical test. • For a given sample size and significance level α, statistical power is evaluated as the proportion of iterations for which $p < \alpha$.	• E.g. for 10000 Monte-Carlo iterations, if $p < \alpha$ for 84% (8400 iterations), Power = 0.84.

2.5. Baseline distributions of each outcome metric

Figure 1 shows a distribution of the LOMV, VFD and LOMV-28 distribution for 10,000 cohorts of 100 patients selected from Cohort A and Cohort B prior to implementing an intervention effect. Patient selection was iterated 10,000 times using random selection with replacement (Table 3,

Step 2) to create the boxplots. As can be seen, both the LoMV and LoMV-28 cohorts have significantly skewed log-normal distributions, whereas VFD shows a reverse log-normal distribution that is highly skewed towards 28 days.

The distribution spikes at the start of the VFD, and end of the LoMV-28 plots are due to the impact of mortality data on these metrics. This clearly shows how any change in mortality due to an intervention can have a further significant effect on the distribution shape. Finally, the LoMV-28 and VFD metrics are also not log-normal given these spikes, which would likely cause further issues when using a trial design method based on a normal distribution assumption, even if the data was logged. Consequently, use of sample size estimation methods that require a Gaussian distribution, which is the common approach, would not have been appropriate for these outcome metrics [17,18], and thus non-parametric statistics and Monte Carlo analysis, as proposed, should provide a more accurate solution.

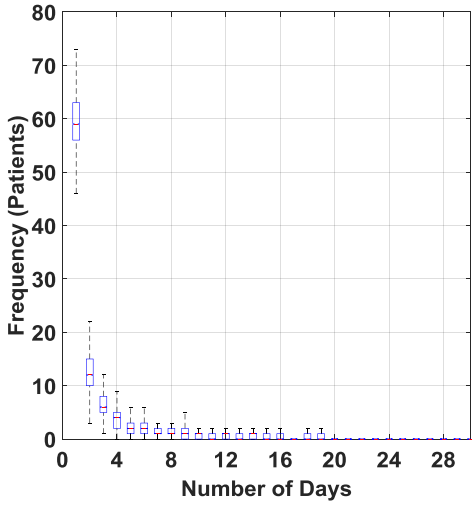
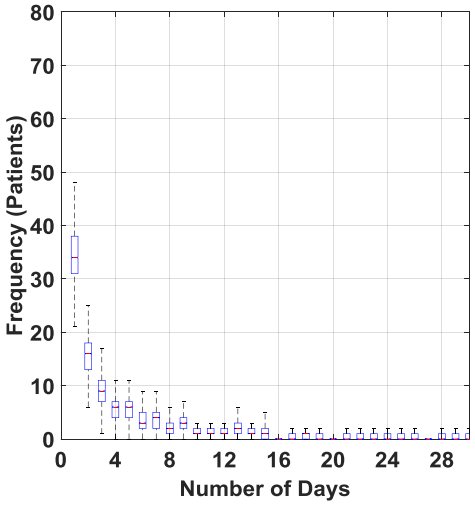
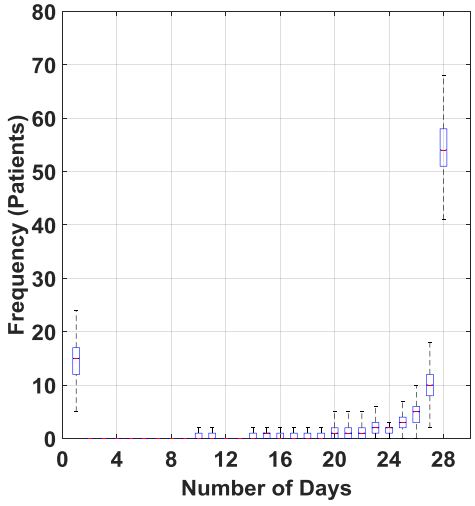
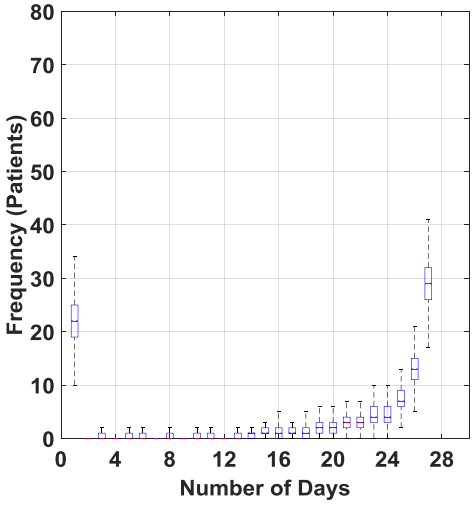
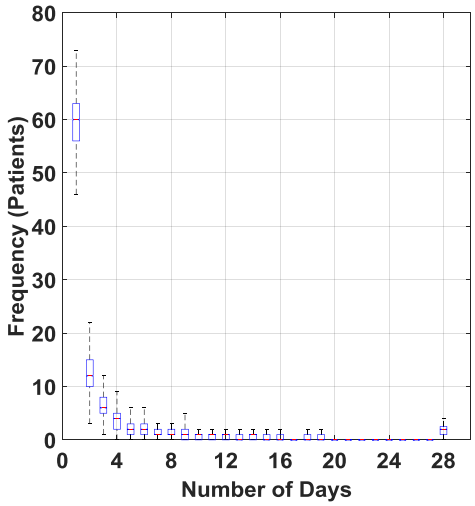
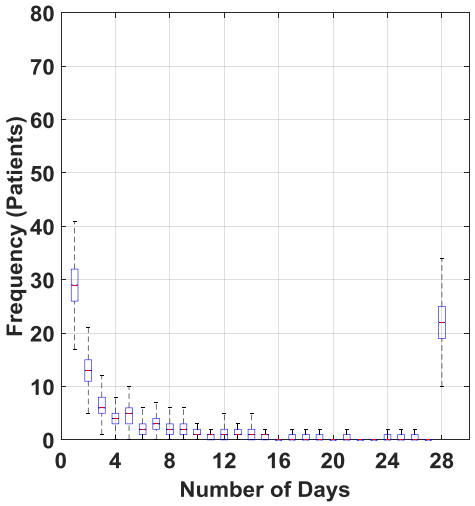
	Cohort A	Cohort B
LoMV		
VFD		
LoMV -28		

Figure 1: Outcome metric distributions for sample size of 100 patients.

2.6. Analyses

Each cohort size for N = 100 – 2000 with a 10 patient step size are analysed as follows:

1. Cohort A (LoMV, VFD, LoMV-28)
2. Cohort B (LoMV, VFD, LoMV-28)

Analyses 1 and 2 use a two-tailed test. Two-tailed tests can separate whether the intervention yields a better or worse outcome. A one-tailed test assumes the intervention is better or not better, but cannot show it is worse. A one-tailed test at $p<0.025$ is considered equivalent to a two-tailed test at $p<0.05$. However, clinically, an intervention that is not better is potentially enough of an answer, as clinicians seek better treatment. Therefore, the impact of single-tail testing in this approach was also considered for:

3. Cohort B (single vs double-tailed tests) for LoMV, VFD, LoMV-28.

Finally, all tests in 1-3 assume equivalent mortality as simulated. However, a good intervention might be expected to reduce mortality, which in turn affects VFD and LoMV-28. This aspect was also simulated, by randomly selecting patients to have their mortality changed in the intervention cohort, and repeating Analysis #2:

4. Cohort B (5% and 10% mortality differential) for the mortality affected VFD, LoMV-28 metrics.

These four analyses clearly delineate the impact on trial design and trial size, using such nonlinear distributions and metrics, for explicit exclusion criteria in cohort selection (Analysis #1 vs Analysis #2), statistical test used (Analysis #3), and the impact of mortality when using mortality affected metrics (Analysis #4).

2.7 Overall Impact of Method, Cohorts and Analyses:

The struggle that the design of many ventilation trials face is the excessive dimensionality of patient factors (diagnosis, age, sex) and MV care factors (how they are treated, and thus the size of the intervention effect). The key to this method and trial design approach is that it collapses that dimensionality in two ways. First, the objective exclusion criteria, eliminate unintended subjectivity and patient dimensions, where subjectively patients may be either included or excluded, creating variability between the trial cohort and the intended target cohort who might benefit. Second, it does so through the use of repeated simulation, thus covering all possible or likely cohort outcomes, where the use of 4-years of data from the trial unit provides a final means of reducing potential unintended variability in this model-based approach. Thus, the use of objective inclusion and exclusion in-silico criteria reduces a lot of dimensionality and uncertainty that would otherwise occur.

The overall non-parametric simulation methods and design approach was selected as it would be feasible in a clinical trial. The objectivity implemented in a manner where it can be used in the actual trial ensures the desired lower dimensionality is preserved. In turn, this outcome provides an increased chance of reaching significance through better control of the trial design and the actual trial so that the trial design is a far better match for what occurs in implementation, increasing the likelihood that if the assumed intervention benefit is observed the trial will be significant.

3. Results

3.1. Sample size estimation for each metric (Analyses #1 and #2)

Graphical results for sample size estimation with a 25% difference in LoMV between control and intervention groups are shown in Figure 2. The X-axis shows the samples size and the Y-axis shows the corresponding power obtained through the 10000 iteration Monte-Carlo simulation at each sample size. Cohort A which included all invasively ventilated patients had a much lower power compared to Cohort B. Due to the negatively skewed distribution of the VFD metric as shown in Figure 1, a log-transformed student t-test was not suitable for the significance testing. This is a point often not noted in the trial design of studies that use VFD.

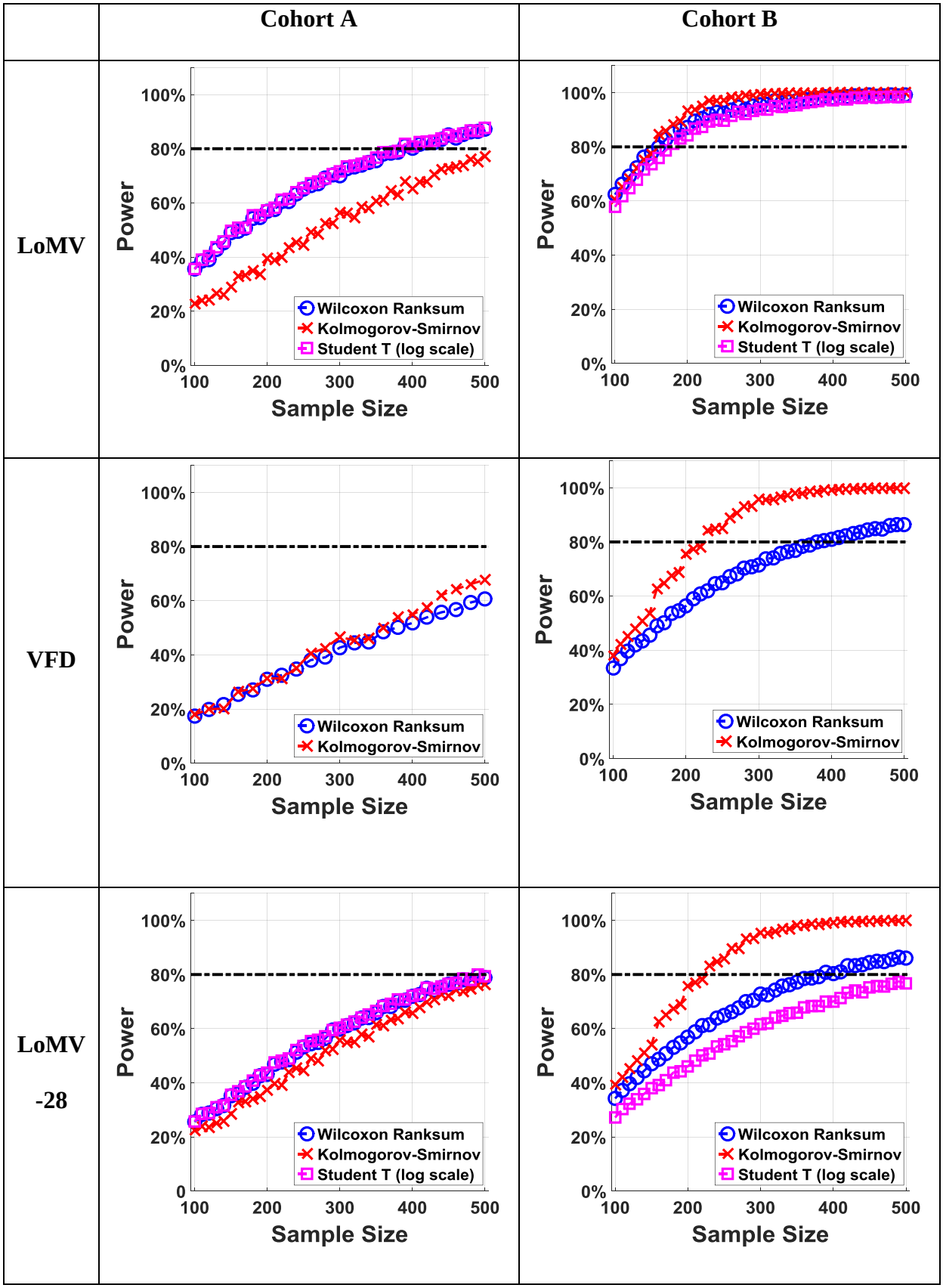


Figure 2. Results of Monte Carlo simulation for 25% LoMV difference

The power analysis was carried out for each metric at LoMV differences (ΔLoMV) of 10%, 15%, 20% and 25% and the estimated sample sizes per trial arm are shown in Table 4. These effect sizes are realistic based on the trials summarised in Table 1.

Table 4: Estimated sample sizes per trial arm for LoMV outcome metric sizes of 10-25% in 5% increments, for both Cohorts A and B, using all 3 statistical tests.

ΔLoMV	**Wilcoxon Ranksum**		**Kolmogorov-Smirnov**		**Student T-Test (log scale)**	
Cohort	**A**	**B**	**A**	**B**	**A**	**B**
10%	2000+	1350	2000+	750	2000+	2000+
15%	1340	530	1530	370	1330	670
20%	670	270	850	240	670	310
25%	400	160	530	160	390	180
ΔVFD						
10%	2000+	2000+	2000+	850	T-Tests were not able to be used for the VFD metric	
15%	2000+	2000+	1800	500		
20%	1460	700	1030	330		
25%	790	390	650	220		
ΔLoMV-28						
10%	2000+	2000+	2000+	860	2000+	2000+
15%	2000+	2000+	860	480+	2000+	2000+
20%	700	700	330	330	1160	1160
25%	490	380	510	230	490	540

Transferring the input parameter uncertainties to ranges is achievable. The uncertainty in patient types and which patients might arrive at a given trial period was covered parametrically by the dimensions N_{cohort}, Δ_{effect} and a reduced mortality on the VFD and LoMV-28 metrics, as well as by using 4-years of patient data from the trial center. Thus, the resulting, Monte Carlo range of p-values from each Monte Carlo run yields the power at a significance level of 0.05, as seen in Figure 3 and Table 4. For clarity, Figure 2 demonstrates the ability of this method to also show the range of p-values that dictate whether or not the sample size can achieve a power of 80%. The box and whisker plots show the results from a 25% LoMV difference, as in Figure 1, but with the

full range of Monte Carlo run results, where Cohort B has the objective trial exclusion criteria applied.

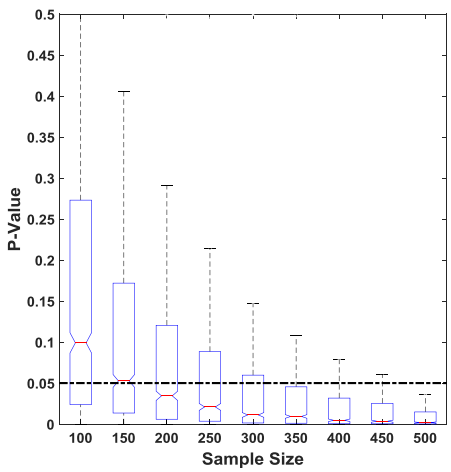
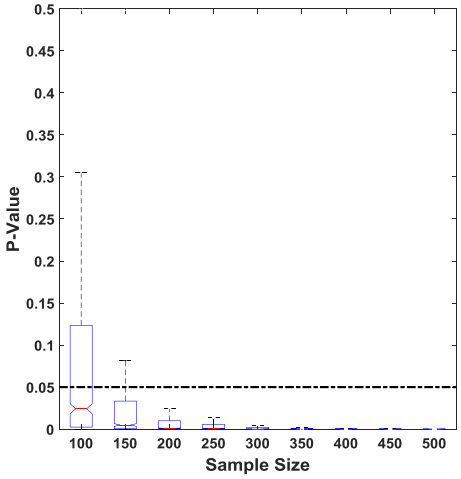
Cohort A	Cohort B
	

Figure 3: Results of each Monte Carlo simulation for 25% LoMV difference in box and whiskers format for both Cohorts A and B, with p < 0.05 shown. The percentage of results below p = 0.05 indicates the power at that sample size.

3.2. Single vs double-tailed tests (Analysis #3)

Single (upper) tailed Wilcoxon Rank-Sum and t-tests were carried out on each metric with a set significant level of 0.05 to assess whether this had an impact on the results. At a LoMV difference of 25%, there is no difference in the required sample size in LoMV and LoMV-28, as shown in Table 5. However, for the VFD metric, a significant reduction of required sample size to achieve 80% power was found.

Table 5: Sample size comparison between single-tailed and double-tailed statistical tests for LoMV, LoMV-28 and VFD when the intervention effect is 25% reduction in LoMV

	LoMV		VFD		LoMV-28	
	Single-Tailed	**Double-Tailed**	**Single-Tailed**	**Double-Tailed**	**Single-Tailed**	**Double-Tailed**
RS-Test	160	160	290	390	380	380
KS-Test	160	160	170	230	230	230
Log T-Test	180	180	N/A	N/A	550	540

3.3. Impact of mortality difference (Analysis #4)

Schoenfeld et al [7] hypothesised that using VFD to determine intervention differences would require a much higher sample size than LoMV if there was not a significant difference in mortality rates. To this end, concomitant mortality rate reductions of 5% and 10% was simulated in the intervention cohort for the LoMV-28 and VFD metrics [2]. The sample sizes for a 25% LoMV difference and 5 to 10% of mortality rate difference for the Wilcoxon-Ranksum and Kolmogorov-Smirnov analyses are shown in Table 6.

Table 6: Impact of simulating a mortality differential between control and intervention cohorts for VFD and LoMV-28 outcome metrics.

	No Mortality Difference		5% Mortality Difference		10% Mortality Difference	
	RS-Test	**KS-Test**	**RS-Test**	**KS-Test**	**RS-Test**	**KS-Test**
LoMV-28	380	230	190	190	110	140
VFD	280	220	130	130	90	100

As can be seen, the sample size required to reach 80% power is significantly reduced if a mortality differential of 5% occurs between each cohort. Such an improvement is reasonably possible for a targeted cohort receiving better care. To capitalise on these findings, it is recommended that the mortality rate is accounted for at the end of the trial to assess the efficacy of the intervention treatment.

4. Discussion

4.1. Impact of intervention effect and exclusion criteria on results

The Monte-Carlo simulation based design method was able to estimate the sample size required for a clinical trial to detect a significant difference with set power. It was found that when the intervention effect is small at 10%, much larger sample sizes of more than 2000 patients per trial arm are required as compared to larger intervention effects, as expected. This larger sample size from the typical design method is due to the skewed and highly variable distribution of ventilation duration of the cohort, and is indicative of the range of patients' underlying conditions that require mechanical ventilation. Importantly, these distributions in Figure 1 are typical and do not match the assumptions made by typical design tools.

When the exclusion criteria are implemented in the simulation, the required sample sizes per arm at the same intervention effect, is further reduced. From Figure 1, it is clear that if Cohort A was considered as the trial cohort, the sample size required for clinical significance and a well-powered study is much higher compared to targeting a specific patient cohort (Cohort B). This result shows that targeting a specific cohort through implementing objective and easily defined inclusion and exclusion criteria available at patient admission can result in a 'narrower' metric distribution, which is important. Thus, a clinical trial that aimed to reduce LoMV, or increase VFD, should be designed to target specific patient groups who are likely to benefit from the treatment and whose distribution of patient-specific LoMV is amended to seeing a change for reasonable sample size [17].

Finally, and importantly, trial sizes also impact on patient risk. A trial with equipoise in its hypothesis includes the risk of the intervention possibly having a negative effect on patient

outcome. Thus, the fewer the number of patients in the trial design that are needed to achieve significance and power, achieved here with a non-parametric Monte Carlo simulation design approach, the lesser the risk to patients in determining the impact and safety of the intervention.

4.2. Impact of different outcome metrics and intervention effect

Sample sizes for different outcome metrics were examined in this study. The LoMV metric was found to have the smallest sample size required to achieve significance with 80% power compared to the other 2 metrics. This was attained with 160 patients required per trial arm for a 25% LoMV reduction difference in Cohort B. However, at a difference of 20% this value increases to 270 patients per arm. 15% and 10% differences see a rise to 1000 patients per arm for an 80% powered study. This finding also shows the perils of these outcome metrics and one possible reason behind non-significant RCTs aside from non-effective clinical interventions. Specifically, if a large intervention effect, which is difficult to achieve, cannot be obtained trial sizes grow rapidly along with the likelihood of other risks to the trial.

Both the VFD and LoMV-28 outcome metrics were studied with the hypothesis that the inclusion of mortality would affect the power of the study. The VFD metric was specifically designed with the intention that a new treatment that either reduced the length of ventilation, or mortality, would be more likely to show a significant difference in a trial [7]. However, this simulation does not include any changes in mortality and hence the effect is minimal and both metrics have led to a lower powered study than the standard LoMV outcome. It is expected that the discrepancy in the t-test results for the LoMV-28 metric is due to the change in distribution shape due to the peak at 28 days. This obstructs the ability of the statistical testing method to detect a significant difference.

4.3. Impact of different statistical tests

Incorrect assumptions about the distribution of data can result in an inconclusive and under-powered study [15]. A two sample unpaired t-test requires data with a Gaussian distribution. The log-normal distribution of the LoMV and LoMV-28 allowed a log-transformed t-test to be carried out on the data. This distribution shape was verified by the similar results from the Wilcoxon Rank-Sum analysis. The negatively skewed VFD metric did not meet the Gaussian distribution assumption for a t-test, even when log-transformed.

The impact of using a single-tailed test with a significance level of 0.025 was assessed for the LoMV and LoMV-28 outcome metrics and found to be minimal. However, it showed a significant improvement in the study power for the VFD metric in Table 5.

It is possible to choose the test with the lowest sample size using this method. This has often been the case for most clinical trials. The criteria for choosing one statistical method over another has often been due to statistical ‘correctness’, a limitation of resources for the trial favouring lower numbers, or due to ethics committee or independent statistician requirements. If it is known that a treatment is not improving prognoses, it is not ethical to continue with it.

Often with skewed data sets, such as LoMV, it is preferable to use a non-parametric statistical test. Hence, of the three tests used in this study (Kolmogorov-Smirnov, Wilcoxon Ranksum and (log) Student T-Test), only the T-Test assumes a distribution., which is then log corrected Ideally, either of the two non-parametric tests would be used, where Kolmogorov-Smirnov is more sensitive to differences in distribution spread, while Wilcoxon Ranksum is more sensitive

to differences in the data set median. Which test should be used would depend on what a given trial is aiming to achieve.

4.4. Absolute vs percentage decrease in intervention effects for VFD outcome metric

As shown in Tables 5 and 6, the VFD outcome metric displayed a considerably lower power than LoMV-28. This result was unexpected as the distributions of each metric are mirrored, and non-parametric tests were used. The inconsistency is due to the use of percentage LoMV reductions. Percentage reductions imply less change for shorter stay patients, which is clinically reasonable vs an absolute change that has lesser impact for longer stay patients. Table 7 demonstrates the discrepancy in percentage changes for an initial LoMV of 5 days, with intervention of 20% LoMV.

Table 7: Differences between absolute and percentage reductions, for initial LoMV of 5 days

	Value Before Intervention	**Value After Intervention**	**Percentage Change**
LoMV	**5**	**4**	20%
LoMV-28	**5**	**4**	20%
VFD	28-5 = **23**	28-4 = **24**	4%

In this case, the VFD metric with a typically seen LoMV of ~5 days means a large change in LoMV and thus LoMV-28 as a percentage is a relatively small change in VFD. At LoMV = 14 days the effect would be equal, and over 14 days the situation would reverse with greater effect for VFD and an easier ability to detect change in this metric. This latter case is shown in Table 8 with an initial LoMV of 20 days, and a percentage reduction of 20%.

Hence, the choice of LoMV-28 or VFD should depend primarily on the given initial distribution, and this outcome would apply more generally to other trial design approaches with similar metrics. In addition, Using LoMV as the root outcome to test in simulation is important, as well as knowing the exact distribution. This Monte-Carlo approach can do this, unlike other commonly used methods.

Table 8: Differences between absolute and percentage reductions, for initial LoMV of 20 days

	Value Before Intervention	Value After Intervention	Percentage Change
LoMV	**20**	**16**	20%
LoMV-28	**20**	**16**	20%
VFD	28-20 = **8**	28-16 = **12**	33%

4.5. Statistical significance and power

A significant problem in many trials is concentrating on the clinical results, while neglecting the statistical significance [9]. Conducting a power analysis allows the probability of correctly detecting a difference between the control and intervention groups to be determined [9]. This process is complicated when highly variable data that is not normally distributed is being analysed. This analysis used Monte Carlo simulation, combined with clinically relevant exclusion criteria as a viable method of determining the power of a study that uses a primary outcome metric of LoMV.

If the specific RCT assessed in this paper solely analysed LoMV, 160 patients in each arm would be sufficient to achieve a statistically significant result with 80% power and an

intervention difference of 25%. However, using an outcome metric that also considers mortality data, such as VFD or LoMV-28, could be beneficial if there is a mortality difference between each cohort. Using the VFD metric with the same intervention effect can reduce the number of patients required to 130/arm with a mortality differential of 5%.

Due to the high variability and skewed distribution of LoMV data, it is a difficult metric to use to assess the power of a clinical trial [15]. However, it remains one of the most effective methods of determining the efficacy of MV treatment. In addition due to the high cost of ventilator therapy, reductions in ventilator duration have significant economic impacts for care units and hospitals [8].

4.6. Limitations

4.6.1. Use of data from a specific ICU

The study was undertaken with the assumption that the data used in the simulation was indicative of that which would be used in the RCT. However, use of LoMV distribution data from a single, specific ICU may mean that the results are not universally applicable, and only able to be used in those with similar characteristics [15]. Nevertheless, the approach followed in this study is general enough to be able to be repeated and utilised for most centres. If information on LoMV distribution is required, then a small and centre-specific pilot study could be carried out. In the case of a multi-centre study, information from multiple centres should be used.

5. Conclusions

In this study, a Monte-Carlo simulation method was able to effectively estimate the sample sizes required for a clinical trial that consists of a highly variable and skewed outcome metric, such as LoMV. The effect of using LoMV, VFD or LoMV-28 was assessed, along with the impact of imposing clinically relevant exclusion criteria. Higher sample sizes for a powered study is required for VFD, as the intervention effect converting from LoMV to VFD is different. However, the use of VFD could be improve if additional mortality rate difference was included.

For a clinical trial, the shape of LoMV distribution is critical, and use of exclusion criteria to target patient groups which may benefit from this intervention is useful. Assessment of ΔLoMV in response to treatment should be considered to avoid an under-powered study. Monte-Carlo simulation, combined with objective patient selection criteria provides better design of ventilation studies. Finally, the overall approach used here is readily generalisable to most trials where the outcome measures are based on a lognormal or otherwise skewed data set, such as most length of care outcomes commonly used in medical research trials.

6. Acknowledgements

The authors would like to acknowledge the funding of the EU FP7 and Royal Society of New Zealand IRSES funding, as well as funding from the NZ MedTech CoRE.

Appendix A

Table A1: Patient number and distribution

	Patient Number				
	2011	**2012**	**2013**	**2014**	**Total**
All patients	1279	1284	1344	1269	5176
MV	1004	953	963	977	3897
Invasive MV	878	825	830	850	3383

Table A2: Patient LoMV distribution

	LoMV distribution (Days) mean ± std, [median, IQR]				
	2011	**2012**	**2013**	**2014**	**2011-2014**
MV	2.54±6.12 [0.62,0.24-1.91]	2.90±6.61 [0.73,0.25-2.20]	2.98±6.49 [0.79,0.24-2.48]	2.88±6.16 [0.83,0.31-2.76]	2.82±6.34 [0.74,0.25-2.29]
Invasive MV	2.66±6.38 [0.62,0.24-1.94]	3.06±6.92 [0.72,0.24-2.32]	3.24±6.86 [0.81,0.24-2.75]	2.88±5.79 [0.80,0.26-2.70]	2.95±6.50 [0.73,0.24-2.48]

Table A3: Mortality LoMV distribution

	ICU Mortality rate (%)				
	2011	**2012**	**2013**	**2014**	**2011-2014**
All patient	8.68%	11.99%	8.93%	10.48%	10.01%
MV	10.66%	15.22%	11.84%	12.59%	12.55%
Invasive MV	11.85%	16.12%	12.65%	13.53%	13.51%

Table A4: APACHE III codes used for exclusion criteria.

	Code	**Description**		**Code**	**Description**
Non - Operative	**100**	Cardiovascular	**Post- Operative**	**1200**	Cardiovascular
	200	Respiratory		**1300**	Respiratory
	300	Gastrointestinal		**1400**	Gastrointestinal
	400	Neurological		**1500**	Neurological
	500	Sepsis		**1600**	Trauma
	600	Trauma		**1700**	Renal/ Genitourinary
	700	Metabolic		**1800**	Gynaecological
	800	Haematology		**1900**	Musculoskeletal
	900	Renal disorder		**2100**	Haematological
	1000	Other medical disorders		**2200**	Metabolic
	1100	Musculoskeletal/ Skin disease			
	0	No diagnosis entered			

*Source from Australian and New Zealand Intensive Care Society (ANZICS), Centre for Outcome and Resource Evaluation (CORE) Adult Patient Database (APD). The details of the APACHE III diagnostic codes can be found in www.anzics.com.au/downloads/doc_download/883-apd-data-dictionary-v4-basic.

Appendix B

Table A5: Exclusion criteria used in clinical trial, and in simulation

Inclusion criteria	
Actual clinical protocol	**Simulation method**
Patients requiring invasive MV Patients with PaO2/ FiO2 (PF ratio) < 300 mmHg Arterial line in situ.	Patients requiring Invasive mechanical ventilation
Exclusion criteria	
Patients who are likely to be discontinued from MV within 24 hours.	Exclude patient with LoMV < 1 days.
Patients with age < 16 years.	Exclude patient with age < 16 years.
Any medical condition associated with a clinical suspicion of raised intracranial pressure and/or a measured intracranial pressure ≥ 20 cmH_2O.	Exclude patient with head trauma using APACHE III diagnostic Code 601 - Head trauma with or without multi trauma 1601 - Post operation patients: head trauma with or without multi trauma
Patients who have a high spinal cord injury with loss of motor function and/ or have significant weakness from any neurological disease.	Exclude patient using APACHE III diagnostic Code 400 - Neurological non-operative 604 - Multi trauma with spinal injury 605 - Isolated cervical spine injury 1500 - Post-operative: Neurological patients 1604 - Post-operation patients: Multi trauma with spinal injury 1605 - Post operation patients: isolated cervical spine injury
Patients who have a barotrauma (pneumothorax, pneumomediastinum, subcutaneous emphysema or any intercostal catheter for the treatment of air leak).	No action performed
Patients who have asthma as the primary presenting condition or a history of significant chronic obstructive pulmonary disease.	Exclude APACHE III diagnostic code 206 - Chronic obstructive pulmonary disease 209 - Asthma
Patients who are moribund and/or not expected to survive for > 72 hours.	No action performed
Patients who have already received MV for > 48 hours (including time spent ventilated in a referring unit).	No action performed
Lack of clinical equipoise by intensive care unit (ICU) medical staff managing the patient.	No action performed

References

[1] A.S. Slutsky, ACCP Consensus Conference - Mechanical Ventilation, Chest. 104 (1993) 1833–1859. doi:10.1378/chest.110.3.866-a.

[2] R.G. Brower, P.N. Lanken, N. Macintyre, M.A. Matthay, A. Morris, M. Ancukiewicz, et al., Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome, N. Engl. J. Med. (2004) 113–116. doi:10.1056/NEJMp1415160.

[3] M.O. Meade, D.J. Cook, Y.M. Arabi, D.J. Cooper, A.R. Davies, L.E. Hand, et al., Ventilation Strategy Using Low Tidal Volumes, Recruitment Maneuvers, and High Positive End-Expiratory Pressure for Acute Lung Injury and Acute Respiratory Distress Syndrome: A Randomized Controlled Trial, Jama. 299 (2008) 637–645. doi:10.1001/jama.299.6.637.

[4] A. Mercat, J.-C.M. Richard, B. Vielle, S. Jaber, D. Osman, J.-L. Diehl, et al., Positive End-Expiratory Pressure Setting in Adults With Acute Lung Injury and Acute Respiratory Distress Syndrome, Jama. 299 (2008) 646. doi:10.1001/jama.299.6.646.

[5] The Acute Respiratory Distress Syndrome Network, • Ventilation With Lower Tidal Volumes As Compared With Traditional Tidal Volumes For Acute Lung Injury And The Acute Respiratory Distress Syndrome, N. Engl. J. Med. (2000) 1301–1308.

[6] D. Breen, T. Churches, F. Hawker, P.J. Torzillo, Acute respiratory failure secondary to chronic obstructive pulmonary disease treated in the intensive care unit: a long term follow up study., Thorax. 57 (2002) 29–33. doi:10.1136/thorax.57.1.29.

[7] D.A. Schoenfeld, G.R. Bernard, Statistical evaluation of ventilator-free days as an efficacy measure in clinical trials of treatments for acute respiratory distress syndrome, 30 (2002).

[8] J.F. Dasta, T.P. McLaughlin, S.H. Mody, C.T. Piech, Daily cost of an intensive care unit day: The contribution of mechanical ventilation*, Crit. Care Med. 33 (2005) 1266–1271. doi:10.1097/01.CCM.0000164543.14619.00.

[9] S.S. Bhardwaj, F. Camacho, A. Derrow, A.B. Fleischer, S.R. Feldman, Statistical Significance and Clinical Relevance, Arch Dermatol. 140 (2004) 1520–1523.

[10] M.-C. Pintado, R. de Pablo, M. Trascasa, J.-M. Milicua, S. Rogero, M. Daguerre, et al., Individualized PEEP setting in subjects with ARDS: a randomized controlled pilot study., Respir. Care. 58 (2013) 1416–23. doi:10.4187/respcare.02068.

[11] T. Strøm, T. Martinussen, P. Toft, A protocol of no sedation for critically ill patients receiving mechanical ventilation: a randomised trial, Lancet. 375 (2010) 475–480. doi:10.1016/S0140-6736(09)62072-9.

[12] C.L. Hodgson, D. V Tuxen, A.R. Davies, M.J. Bailey, A.M. Higgins, A.E. Holland, et al., A randomised controlled trial of an open lung strategy with staircase recruitment, titrated PEEP and targeted low airway pressures in patients with acute respiratory distress syndrome., Crit. Care. 15 (2011) R133. doi:10.1186/cc10249.

[13] M.B.P. Amato, M.O. Meade, A.S. Slutsky, L. Brochard, E.L.V. Costa, D. a. Schoenfeld, et al., Driving Pressure and Survival in the Acute Respiratory Distress Syndrome, N. Engl. J. Med. 372 (2015) 747–755. doi:10.1056/NEJMsa1410639.

[14] C.A. Blum, N. Nigro, M. Briel, P. Schuetz, E. Ullmer, I. Suter-Widmer, et al., Adjunct

prednisone therapy for patients with community-acquired pneumonia: a multicentre, double-blind, randomised, placebo-controlled trial., Lancet. (2015) 1511–1518. doi:10.1016/S0140-6736(14)62447-8.

[15] J.H. Van Der Lee, M.W. Tanck, J. Wesseling, M. Offringa, Pitfalls in the design and analysis of paediatric clinical trials: a case of a "failed" multi-centre study, and potential solutions, Acta Paediatr. 98 (2008) 385–391. doi:10.1111/j.1651-2227.2008.01048.x.

[16] D.G. Altman, J.M. Bland, Absence of evidence is not evidence of absence., BMJ. 311 (1995) 485. doi:10.1136/bmj.311.7003.485.

[17] Y.S. Chiew, C. Pretty, E. Moltchanova, C. Scarrott, D. Redmond, G.M. Shaw, et al., Reducing the Length of Mechanical Ventilation with Significance : A Case Study of Sample Size Estimation Trial Design Using Monte-Carlo Simulation, in: 9th IFAC Symp. Biol. Med. Syst. 31 Aug-2 Sep 2015, Berlin, 2015.

[18] A. Petrie, J.S. Bulman, J.F. Osborn, Further statistics in dentistry. Part 4: Clinical trials 2., Br. Dent. J. 193 (2002) 557–561. doi:10.1038/sj.bdj.4801627.

[19] D.B. Richardson, Power Calculations for Survival Analyses Via Monte Carlo Estimation, Am. J. Ind. Med. 44 (2003) 532–539. doi:10.1002/ajim.10310.

[20] V.M. Gastañaga, C.E. McLaren, R.J. Delfino, Power calculations for generalized linear models in observational longitudinal studies: A simulation approach in SAS, Comput. Methods Programs Biomed. 84 (2006) 27–33. doi:10.1016/j.cmpb.2006.07.011.

[21] MATLAB R2014a, (2014).